The Journal of Teaching Language Skills (JTLS)
5 (2), Summer 2013, Ser. 71/4
ISSN: 2008-8191. pp. 27-50

A Corpus-Based Study of the Lexical Make-up of Applied Linguistics Article Abstracts

H. Farjami
Assistant Professor, TEFL
Semnan University, Semnan
email: hfarjami@profs.semnan.ac.ir

Abstract

This paper reports results from a corpus-based study that explored the frequency of words in the abstracts of applied linguistics journal articles. The abstracts of major articles in leading applied linguistics journals, published since 2005 up to November 2011 were analyzed using software modules from the Compleat Lexical Tutor. The output includes a list of the most frequent content words, lists of frequent words and abbreviations not found in the British National Corpus. The study also weighed applied linguistics abstracts against the General Service List and the Academic Word List and identified words in these abstracts which are shared by the GSL or the AWL or are unique to one set. The report separately lists words from the GSL and the AWL which are proportionally more frequent in these abstracts than in general written texts, and hence may be reasonably regarded as playing key textual roles in applied linguistics abstracts and, by extension, discourse.

Keywords: abstracts, applied linguistics, AWL, frequency, GSL, lexical make-up

1. Introduction

Surveying journal abstracts seems to provide a practical and valid reservoir of condensed information. They are practical for reasons of ready availability and terseness and valid because, this genre, as Swales and Feak (2009) suggest, shows best the features of specialized communication between experts in the related field. Moreover, research article (RA) abstracts are expected to represent the issues raised and the ideas discussed in the body and describe accurately and briefly the contents of the whole text

Received: 10/21/2012 Accepted: 5/21/2013

(Lores, 2004). This means that a comprehensive analysis of RA abstracts published in journals which represent a field can provide an outline of the developments in that field. Similarly, analyzing the abstracts of the journals related to applied linguistics may furnish useful insight into targeted themes. In fact, it was due to understanding this potential that the study reported here was conducted. The report presents a corpus-based survey of the more frequently-used words in applied linguistics article abstracts (ALAAs) published since the beginning of 2006 up to November 1, 2011with the hope that the lexical profile which emerges will give the interested scholars and future authors better orientation, add to the professional awareness of researchers and practitioners, and provide handy word lists for practical use.

2. Studies of Abstracts

An abstract, according to Bhatia (1993), is 'a description or factual summary of a much longer report, and is meant to give the reader an exact and concise knowledge of the full article' (p. 78). According to the APA manual (American Psychological Association, 2010), a good abstract should be accurate, self-contained, concise and specific, non-evaluative, and coherent and readable.

It is generally acknowledged that research article RA abstracts play a key role in the academic and scientific sphere around the world. In fact, as Ventola (1994) stated, abstracts "have become tools of mastering and managing the ever increasing information flow in the scientific community" (p. 333). So, quite a number of authors have studied RA abstracts as well as their variations across disciplines (e.g. Huckin, 2001; Hyland, 2004; Martin, 2003; Samraj, 2005). It has been established that RA abstracts differ from the main body of articles in their lexical, thematic and rhetorical structure and constitute a genre in their own right although the two, not surprisingly, share many features.

These studies have approached the problem from varying perspectives. Many of them have delineated the function and macro-organization of the abstracts in the targeted genres (e.g. Hyland, 2004). Others have focused on the lexico-grammatical features of the abstracts to give an in-depth picture of one or two linguistic features of the abstract. One important aspect is the move structure. Abstracts are made up of moves, which can be characterized, according to Lorés (2004), as a "functional term that refers to a defined and bounded communicative act that is designed to contribute to one main communicative objective, that of the whole text" (p. 282). Some researchers have turned their attention to flaws in abstracts (e.g. Salager-Meyer, 1990). Other approaches to the study of abstract include inter-lingual comparative studies of abstracts (Martin, 2003), the rhetorical structure of

abstracts (e.g. Hartley & Sydes, 1997), and the organization of themes and rhemes (e.g. Ghadessy, 1999).

Some studies investigate abstracts from broad areas such as humanities, social sciences and natural sciences. Other studies examine abstracts in a specific discipline. For example, Busch-Lauer (1995), and Salager-Meyer (1990) focused on the discipline of medicine; Huckin (2001) on biomedicine; and Hartley (2003) on psychology. The abstracts of the articles in the field of applied linguistics, the discipline of interest in the present study, have also received some attention (e.g. Hyland, 2004; Lorés, 2004; Pho, 2008; Santos 1996). Hyland's (2004) study compared the move structure of abstracts across eight disciplines; one of those disciplines was applied linguistics. Santos (1996) was probably the first endeavor to establish the textual organization of ALAAs. Focusing exclusively on the field of applied linguistics, Santos (1996) selected 94 abstracts of applied linguistics articles to study and found a prevalent five-move model with sub-moves. Santos also examined the distribution of a few linguistic features such as verb tenses across moves. Lorés (2004) and Pho (2008) were two small-scale studies, focusing, respectively, on the thematic organization and authorial stance of abstracts.

Not surprisingly, studies of RA abstracts have been overwhelmingly corpus-based. Corpus-based methodology, which has considerably increased over the last three decades due to improvements in computer techniques and recognition of the value of large-scale corpora in studying actual language use, is frequently applied to the study of different aspects of lexis. In quantitative corpus studies, researchers identify and classify particular lexical patterns, count them, evaluate them statistically and sometimes develop models to explain what is observed (McEney &Wilson, 2001). An example of quantitative corpus research which is very similar in aim and method to the present one is the study by Vongpumivitch, Huang, and Chang (2009), who explored the use of words in Academic Word List (Coxhead, 2000) in the field of applied linguistics. The result of their analysis was the identification of the frequency and range of AWL and non-AWL word forms across five applied linguistics journals.

3. Value of Abstract Studies

Languages show preference for particular rhetorical and linguistic strategies, observed in the distribution and frequency of certain structural, semantic and pragmatic patterns. Technical writers and experts should adapt their abstracts to the features of the English language for the particular communicative setting; otherwise, the target discourse community may reject those texts because they do not comply with the expectations readers

have. Awareness of, and understanding, the micro- and macro-linguistic patterns favored by the RA abstracts of particular fields seem to be essential for those who aspire that their research and findings can be successfully reported and accepted by other members of the discourse community (see Groom, 2005). Similarly, certain terms and phraseological units in a given field tend to be used more frequently; therefore, observing the established norms and patterns of frequency on the part of authors, too, contributes to their writing to the genre and, hence, bears a part in successful expert-to-expert communication (Chan & Foo, 2004).

4. Word Lists

There have been longstanding attempts to identify the more frequent words specific to academic discourse and to determine their frequency profiles. One of the first such attempts was the compilation of a General Service List (GSL) by West (1953). It contains the 2000 most widely and frequently used English word families from a corpus of five million words. This list has had a wide influence through the years, serving as the basis for graded readers as well as other material. Although developed 60 years ago, the GSL covers up to 90% of fiction texts, up to 75% of non-fiction texts, and up to 76% of academic English (Coxhead, 1998). Bauman and Culligan (1995) modified the GSL and ended up with 2284 head words.

There have also been several attempts at compiling lists of the most frequent and/or useful academic words. Praninskas (1972) compiled corpora-based lists of words which occurred across a range of texts. Xue and Nation (1984) combined and edited four existing lists and developed the well-known University Word List (UWL). More recently Coxhead (2000) compiled another well-known list— the Academic Word List (AWL) – from a corpus of 3.5 million written academic words outside the first 2000 most frequent English words (GSL). Coxhead (2000) emphasized the fact that the 570 items in the AWL covers about 10% of tokens in academic passages but only 1.4% of the tokens in fictional texts as proof that the list contains predominantly academic words. The AWL has, in return, been a reference and point of departure for many EAP vocabulary textbooks and exercises and continues to encourage and inspire further research. For example, Simson-Vlachh and Ellis (2010), inspired by the AWL, developed a corpus-based Academic Formula List (AFL), which includes formulaic sequences of words frequently recurring in academic written and spoken discourse. The AWL has been validated by some researchers. For example, Vongpumivitch et al. (2009) did a frequency analysis of the words in applied linguistics research papers and found that the AWL accounted for 11.17% of their

corpus, which was slightly higher than the coverage reported by Coxhead (2000) for academic texts in general.

The rationale for specifying the lexical make-up of target discourse domains for instruction is simple. According to nation (2001), the more frequent items have the highest utility and should therefore be taught earlier than the less frequent ones. A similar logic can be employed in justifying researching the lexis of article abstracts, although RA abstract may not be typically used for instruction in basic courses in the ESP sense of the word, i.e., for bringing about vocabulary and idiomatic knowledge. Specifying the frequencies of the linguistic features of English RA abstracts including ALAAs seems to contribute to the socialization and initiation of people into their target discourse community, and enhance genre and language awareness. We can speculate that identifying the most frequent content words in ALAAs can, for one thing, considerably optimize their production. This may be, incidentally, in keeping with Ventola (1994), who complained about a lack of useful advice on how to write comprehensible abstracts and Pho (2008), who complained that the current handbooks on research papers either do not mention how to write an abstract at all or only give a general description of an abstract.

5. The Study

The goal of the present study was to examine the frequency of the words used in the abstracts of applied linguistics articles. The investigation is based on a corpus of ALAAs pooled from all the major articles in fifteen applied linguistics journals published since 2005 up to November 1, 2011. The study also aspired to compare the frequency of ALAA lexicon with word lists compiled based on other domains of general and specialized communication. The assumption was that such analysis would help refine the way this field of study is mapped in the minds of the people concerned, as abstracts, by definition, are supposed to be the textual artifact most representative of the ideas being circulated in an academic field. The initial general question which guided the study was: *Which words other than the function words occur frequently in ALAAs?* However, this rough question needed to be further fine-tuned in reference to already established general and academic lists to help obtain a clearer picture of the ALAAs lexical make-up. Hence, the following working questions were formulated to achieve the goals of this study:

1. *What are the 100 most frequent content words in ALAAs?*
2. *What is the share of the GSL and the AWL in the 100 most frequent content words in ALAAs?*

3. What are the most frequent ALAA-specific words and abbreviations?
4. What words play a key role in ALAAs?
4. a. Which GSL words have a key role in ALAAs? That is, which GSL words occur proportionally more frequently in ALAAs than in general reference corpora?
4. b. Which AWL words have a key role in ALAAs? That is, which AWL words occur proportionally more frequently in ALAAs than in general reference corpora?
5. How are the frequency ranks of AWL words different from their ranks in ALAAs?

5.1 The corpus

This study was interested in the lexical profile of the RA abstracts of leading international academic journals in the field of applied linguistics published in 2006 up until the first of November 2011. The list of 157 "linguistic journals" in the Social Science Citation Index from the Thomson Reuters Master Journal list was used because these journals are published by leading international academic publishers and have relatively high impact factors (Aalst, 2010). A shorter list of journals was made based on the titles and, in some cases, after reviewing the descriptors and the contents tables for the journals. The short list of 50 journals was circulated around to colleagues and Ph.D. students well into the field of English language teaching. They were asked to mark the top 15 journals for their association with language teaching, whether the association was general or with a specialized subfield, e.g. assessment and testing. In keeping with their feedback, 15 journals which received the highest additive ranking scores were selected. Not surprisingly, it turned out that many of the selected journals were among the high impact-factor and top-ranking journals identified by Google Scholar, as reported by Aalst (2010), e.g. *Applied Linguistics*, which enjoys a 5-year impact factor of 2.068, *Modern Language Journal*, *TESOL Quarterly*, and *Journal of Second Language Writing* (See Appendix for the list of journals).

Using the copy-and-paste procedure, the researcher collected the abstracts from the journals websites for convenience, as they are the same as those in the print versions of the periodicals. Then, words other than those in the titles and the body, e.g. authors, affiliations, publishers' information, were pruned from the abstracts. This produced an electronic corpus of 2071 abstracts in Microsoft Word format including 377,378 words (See Appendix for statistics for each journal)**.** The fifteen files were amassed and the larger collection of abstracts was then reviewed several times to remove misspellings, using the Find function and the WordPerfect spellchecker set on the American-English mode, although the software used allowed for

spelling variation. The Word file was converted to .txt format to make it compatible with the intended data analysis software.

5.2 Data analysis software

The software modules employed in this study were obtained from the Compleat Lexical Tutor, version 6.2 (Cobb, 2011), which is a web-based suite of lexical analysis tools freely available at www.lextutor.ca, for the purposes of vocabulary teaching and research. The pieces of software used included Familizer Proto, version.5, KeyWords Extractor, version 1, Text Lex Compare, version 2.2, Web Frequency Indexer, version 1.3, and The Compleat Lister, version 2.3. The reason the researcher decided to feature this package in his analysis was lack of access to commercial software, such as WordSmith Tools and MonoConc Pro, which require a license.

The well-known word lists used in comparisons, either by the researcher or underlying the software tools, included Academic Word List (570 words, Coxhead, 2000), General Service List (2284 words, Bauman and Culligan 1995), Brown Corpus list (based on one million words, Francis and Kucera, 1982), and the British National Corpus (BNC) list, (based on one hundred million words, 2007).

6. Results and Discussions

In this section, first, a general overview of the quantitative features of ALAA lexis and of the ALA word list which emerged is presented. Then, the ALAA word list is compared with the GSL and the AWL. Due to space constraints, only the more significant portions of the output from the analyses are presented and discussed**.**

6.1 A general picture

The whole list of the words extracted from ALAAs included 377,378 tokens, and 15,763 types. 5,346 families or lemmas were within the BNC 20,000 words; but 3,593 word types could not be lemmatized as they were outside the BNC, the reference list of the software used, i.e. Familizer Proto v. 5. These included uncommon proper names, acronyms, abbreviations, unconventional numbers and other strings. *The*, *of*, *and*, *in, to,* and *a* stood at the top of the frequency list of ALAA words. This line-up is a little different from the order in well-known general corpora (e.g. Brown Corpus: *the, of, and, to, a, in*; Cobuild General Corpus: *the, of, and, to, a, in*; the BNC: *the, of, and, a, in, to*; the GSL: *the, be, of, and, a, to*). The arrangement is also a little different from the six most frequent words in the specialized academic corpus created by Flowerdew (2001): *the, of, and, to, a, in,* which is the same as Brown Corpus.

The consistency or inconsistency of the rankings of function words across different corpora, however slight, must have to do with generic factors including both semantic and grammatical ones. Compelling evidence for the reality of genre and its lexico-grammatical manifestations comes from the fact that when the most frequent ALAA words in the corpus are separately checked against words in each of the 15 constituent banks of abstracts, they show the *same* rankings up to the 10th most frequent words in all of them. Subsequent items show strikingly similar rankings across the 15 banks, too. This consistency is not confined to function words. Content words also show remarkable consistency in frequency ranking across these applied linguistics journals. *Language* holds the 7th rank both in the main corpus and in all the 15 subcorpora. Other top ranking content words of the corpus hold either the same or very similar positions in the individual subcorpora. This can be yet another indication of the fact that there is not a clear-cut distinction between lexis and grammar but they are inseparable and closely associated as is extensively discussed by Romers (2009). Although the identification of factors which give rise to the frequency patterns of function and content items in different text-types is a worthwhile endeavor in its own right, the general point here is that similar contextual, contextual, and semantic forces bring about similar function words and generic terms and thus help genre specific patterns emerge.

1,334 words in the corpus belong to the GSL, 950 GSL word families do not occur in the corpus, 543 belong in the AWL, 27 AWL words do not occur in ALAAs. Table 1 provides the 100 most frequent content word families along with their frequency tags. Following Coxhead (2000), this list includes lemmas not types.

Table 1. The 100 most frequent content word families in ALAAs*

discuss 760	instruct 926	develop 1215	language 5344**
practice 745	process 911	speak 1173	learn 4618
investigate726	word 833	article 1167	teach 3077
suggest 723	context 811	task 1151	study 2808
show 716	examine 795	find 1069	student 2705
data 705	relation 793	difference 1063	use 2531
text 699	level 786	group 1053	English 2387
foreign 692	assess 782	base 1036	write 1988
proficient 686	acquire 772	read 1036	research 1723
present 672	know 769	result 1032	second 1445
educate 670	participate766	effect 1021	analyze 1345
paper 663	interact 764	room 999	test 1303

type 439	pedagogy 468	first 547	strategy 653
EFL 438	model 466	academy 538	linguistic 634
way 436	significant466	course 538	school 632
need 435	class 464	theory 518	focus 627
describe 430	construct 464	make 514	identify 625
role 424	measure 456	compare 508	form 596
indicate 422	self 456	culture 507	approach 590
able 417	experience 454	vocabulary 507	report 586
argue 415	work 454	meaning 499	provide 585
time 407	vary 451	program 495	native 584
comprehend 403	discourse 448	Spain 490	perform 574
feedback 402	specific 443	explore 474	communicate 562
complex 394	university 443	understand 474	high 552

*Parts of compound words are also included in the counts.
**The figures here and other tables are raw counts unless otherwise indicated.

Not surprisingly, *language*, which holds the 7th rank in the list of all words including function words, is the most frequent content word (5,344 out of 377,378) and accounts for 1.41% of the tokens. The next seven most frequent words (*learn, teach, study, student, use, English,* and *write*) are mainly language and literacy-related specific words. Obviously, one strong source of text identity and belonging to this particular text-type is provided by using these and other frequent content words. Having a probabilistic text schema close to these counts can help the readers easily relate to ALAAs and authors possessing such a quantitative schema are more likely to write and produce characteristically applied linguistics RA abstracts.

Reflection over a descending frequency list of words sampled from a specialized field will tell us what ideas are active in that field. *Write, speak, and read* feature in Table 1 because research on these skills tops the agenda in applied linguistics, while *listen*, ranking 139th with a frequency of 332, has received less attention and falls outside this list. Research-related terms such as *research, analyze, article, result, effect, participate, data,* and *investigate* hold a large share in this table because ALAAs are concerned very much with research methodology. In fact, such tabulation of words can be an ideational map of a discipline if it is based on sound sampling and may provide an opportunity to compare the ideational make-up of different disciplines. An interesting fact, which is also likely for other similar lists and frequency analyses, is that as one moves down this frequency-ordered list, the frequency distance between the adjacent words diminishes. For example, *teach* (3th) and *study* (4th) are 269 occurrences apart, while *comprehend*

(144th) and *feedback* (145th) are only one occurrence apart. When one reaches words with a frequency as low as 30 in the frequency-ordered list of all words, the frequency distance disappears for a sequence of eighteen words.

Of the 100 most frequent content words families, 35 are also very frequent in general texts and therefore are not calculated as "key". However, the other 65 play a more significant role in ALAAs than in general texts and therefore are calculated as ALAA keywords (see below). Sixty-two words families are shared by the GSL and 23 are in the AWL. Only 15 are not included in these two lists. The 15 words which are neither in the GSL nor the AWL are displayed below:

comprehend	experience	meaning	significant
construct	feedback	participate	Spain
discourse	investigate	pedagogy	vocabulary
EFL	linguistic	proficient	

The ALAA words were also analyzed to identify the words which are particularly active in this text type. So, a list of "keywords" was created, using KeyWord Extractor v. 1. This program KeyWord Extractor v. 1 determines the defining lexis in a specialized corpus, by comparing frequency per word to frequency in Brown Corpus as a reference composed of 500 written texts of more than 2000 words on a broad range of topics. The words identified here as key are the word types in ALAAs which are proportionally far more frequent than they are in the Brown Corpus. In this analysis, all the words in ALAA sample at least 10 times more numerous than in the Brown Corpus were identified as key. For example, the first item in the output, *proficiency*, was calculated on the basis that *proficiency* has 3 natural occurrences in the Brown's One million words, but 636 occurrences in ALAA 377,378-word text. These 636 occurrences are proportionally far more numerous in ALAAs than the 3 occurrences in the Brown Corpus. Likewise, the word family, *prime*, with a frequency of 33 is counted as key, while the word *use*, with a frequency of 2531is not.

Thirteen of the 503 key types are proper nouns and their derived adjectives, topped by *Spanish* and followed by *English, Korean, Portuguese, Arabic, Hong Kong, Brazilian, Chinese, Dutch, Iran, Japanese, Taiwan,* and *New Zealand*. Their appearance in the list of keywords is interesting and provides useful information about their referents, although one should remember that proper words, due to being usually infrequent in the reference corpus, have a higher chance of gaining keyness.

The program KeyWord Extractor v. 1 does not currently handle word families and only identifies keyword types along with keyness scores representing the times they are more frequent in, say, ALAAs than in the Brown Corpus. Five-hundred and three word types (375 families) in ALAAs reached the keyness threshold. To short list these types, 340 key types with the highest keyness score were lemmatized with the keyness scores of family members added together. This procedure put out 270 families. The 100 word families with the highest keyness score is shown in Table 2 in alphabetical order.

Table 2. The list of 100 word families with the highest keyness in ALAAs

acquire	educate	interpret	proficient
analyze	English	interview	prompt
aptitude	environment	investigate	qualitative
article	examine	journal	questionnair
assess	expert	language	e
centre	explore	learn	random
challenge	facilitate	lexical	receptive
classroom	feedback	linguistic	rely
cognitive	find	mainstream	research
communicate	focus	metaphor	score
complex	framework	method	self
comprehend	gender	morphology	semantic
compute	genre	motive	speak
concept	globe	narrate	strategy
conclude	grammar	notice	synchronous
construct	grammatical	noun	task
context	highlight	novice	teach
corpus	hypothesis	oral	text
correct	identify	participate	transcript
correlate	immerse	peer	vary
curriculum	implicate	perception	verb
digital	incorporate	perspective	video
discipline	instruct	phonetic	vocabulary
discourse	integrate	pragmatic	
discuss	interact	prime	
domain	interface	problem	

However, as Stubbs (2010) maintains, “keywords are the tips of icebergs: pointers to complex lexical objects which represent the shared beliefs and values of a culture” (p.23). If we are to do more than scratching

the surface and get insights into the beliefs and values which give rise to the keywords listed here, we need to examine them in the context they occur, bearing in mind all the levels of meaning of keyness. Here, keyness is used in a purely statistical sense and fails to relate to levels of culture and schema, nor does the paper explain the relation of the listed "keywords" with other words within phrases (see Stubbs, 2010).

Examining and awareness about exclusively ALAA words and the similarities and differences in patterns and frequencies of words shared by other established lists can also be of significance to the users and producers of ALAAs and enhance their textual schemata. So, more comparisons were made between ALAAs on the one hand, and the AWL, the GSL, and the BNC Corpus on the other. Portions of the output of these comparisons are presented below. Table 3 displays the words with more than ten occurrences in ALAAs but not found in the BNC 20,000 words. Table 4 shows the abbreviations with more than ten occurrences in ALAAs but not included in BNC list. The comparisons with the GSL and the AWL are presented and discussed in the following sections.

Table 3 shows the 69 most frequent words unique to ALAAs. It became possible because the program Familizer Proto, which is based on the BNC and lists lemmas of words used in texts, also lists types which it cannot lemmatize because they are outside the BNC and unique to the text under analysis. The list of words unique to ALAAs was meticulously checked against the original alphabetical and frequency lists of ALAA words using the Find function in the Office Suite to make sure thatcompounds, hyphenated compounds, or other variations, are also considered. Words that existed in both hyphenated and non-hyphenated forms were check with the BNC list to make sure they were not in it in alternative forms. Some of these ALAA-specific top-ranking word types were manually allocated to lemmas, with their frequencies added up. Some words which occurred only in one derived form, e.g. *misspellings* and *codeswitching*, or the derived forms of which seemed more central to ALAAs, e.g. multidimensional, were not assigned to lemmas. The list was, then, curtailed to the items with more than ten occurrences. A large proportion was crosschecked using the Find function in the amassed Word file to make sure of the precision of the final frequencies reported here. Many of the items in Table 3 are proper nouns-- *Hong Kong* being the most frequent, and *Vygotsky* the first personal name to appear in the list-- or compound words, whose components are found in general corpora, e.g. *sociocultural* and *metacognitive*. Some items are not unique to ALAAs in their base forms but the types used in ALAAs do not feature in the first 20,000 frequently-used words in the BNC, e.g. *inferencing* and *processibility*. The most frequent word unique to ALAAs is

nonnative, whose most frequent form, *non-native,* occurs 76 times; hyphenless *nonnative* occurs as frequently as *reflect, assess, constraints, describe, determine, error, example, finally, improve, intercultural,* and *others* (f, 75). The 69th content word, *videoconferencing*, has a frequency of 10 and co-ranks with *varies, videos, syllabi* and 162 other BNC types.

Table 3. The 69 most frequent words outside the 20,000 BNC Words

nonnative 179	morphosyntactic 33	Lardiere 18	generalizability 14
posttest 117	videotape 31	washback 18	Swales 14
sociocultural 101	pretask 31	misspellings 17	Ellis 13
Hong Kong 90	Vygotsky 29	phraseological 17	Catalan 12
intercultural 77	Wagner 28	argumentation 16	offline 12
metalinguistic 76	comprehensibility 26	confirmatory 16	perfective 12
lingua franca 72	affordance 26	memorization 16	Cantonese 11
metacognitive 65	Singapore 26	metacognition 16	dictogloss 11
blogging 60	dialogic 25	languaging 16	expository 11
interlanguage 50	inferencing 24	sociopragmatic 16	Halliday 11
email 50	inferential 24	postsecondary 16	imperfective 11
multimodal 47	Rasch 24	dependability 15	transformative 11
pretest 44	processability 23	multicompetence 15	Horwitz 10
monolingual 39	intertextuality 23	practicum 15	uninterpretable 10
wiki 39	multidimensional 22	metalanguage 15	videoconferencing 10
examinee 36	Cambridge 20	Robinson 15	
crosslinguistic 35	codeswitching 20	prototypical 15	
internet 33	clitics 20	Dornyei 14	

It is worth noting that some words are repeated several times in only a few or even one abstract and, therefore, should not be taken as playing an overall key role as is the case with *misspellings*, whose 17 occurrences are in one abstract. But these are not many and this caution should be applied only to the low-ranking items.

Originally, there were 148 non-BNC items with frequencies above ten. Of these, 69 were content words as presented in Table 3, and 79 were abbreviations (Table 4).

Table 4. The 79 most frequent abbreviations not in the 20,000 NBC

ESL 338	NS 44	DE 28	VIS 20
SLA 181	HL 43	LRES 26	WCF 20
FL 113	CMC 38	CLT 25	DA 19
ELT 89	ELLS 35	RAS 23	EU 19
EAP 87	TOEFL 33	SCMC 22	NNES 19
WH 69	CLIL 32	III 20	CEFR 18
TESOL 67	II 31	LS 20	EI 18
CA 58	WTC 31	NSS 20	RR 18
NNS 47	SA 29	UG 20	ACTFL 17

ERP 16	RA 14	AFL 12	ZPD 11
ESOL 16	SBA 14	WM 12	DST 10
DIF 15	TBLT 14	FFI 11	ELD 10
ICT 15	ANOVA 13	FLES 11	ELL 10
NI 15	CAF 13	IBT 11	ELP 10
OPI 15	CBT 13	IMGS 11	IRF 10
PBL 15	CDA 13	ITAS 11	IWB 10
ASL 14	IPA 13	RP 11	LA 10
IRT 14	PI 13	TLD 11	NNSS 10
NCLB 14	SOPI 13	TOEIC 11	TESL 10
NES 14	TE 13	USA 11	

6.2 Comparing ALAA list with the GSL

1,334 word families in ALAAs are shared by the GSL, which means that 950 GSL words do not occur in these abstracts. Apart from the shared function words, which tend to be the most frequent in virtually all texts, 60 GSL words figure in the top 100 ALAA content words.

The fact that 60 words in the top 100 ALAA words belong to the GSL is interesting because an abstract by definition includes highly condensed language and expectations may be high that it includes relatively fewer general terms than mainstream texts. It takes comparative text studies to pass a judgment as to the comparative lexical density (the ratio of idea units to lexical units) of abstracts and other texts, but the dominance of general words in ALAAs may warrant us to think of other sources of compactness for abstracts than specialized words. One implication can be that complexity of ideation is not always due to prefabricated complex technical words. Compactness and complexity are also created in the combination and interaction of words. So, it is possible to explain complicate ideas by simple and elementary words, at least to some extent. This implication is supported when we examine the list of 503 ALAA keywords and compare it with those shared by the GSL (Table 5). Ninety-five (25.33%) of the keywords in ALAAs belong to the GSL, which means that they are more frequent in ALAAs than in general texts. As an examination of these words can shed light on the current mainstream concepts in applied linguistics and help develop a lexical profile of ALAAs, these words are presented in Table 5 below.

Table 5. 95 GSL words occurring as ALAA keywords

able	adopt	article	centre
accept	advance	base	compare
account	aim	begin	compose
accountable	apply	behavior	content
add	argue	bundle	converse

correct	influence	place	sentence
critic	inform	practice	set
describe	inquire	prefer	skill
develop	introduce	present	speak
difference	know	problem	spell
discipline	language	prompt	standard
discuss	learn	propose	strength
educate	lesson	quantity	student
effect	level	rate	study
English	listen	read	suggest
examine	model	reflect	teach
exchange	native	represent	test
explore	notice	result	track
find	noun	review	translate
foreign	outline	room	use
frequent	pair	scale	verb
gap	paper	school	vowel
grammar	pattern	second	write
include	pause	self	

6.3 Comparing ALAAs with the AWL

Coxhead's (2000) development of the AWL is considered the most significant recent development in the quantitative investigation of academic vocabulary profile (Simpson-Vlach & Ellis, 2010). His application of frequency and range of distribution to a corpus of 3.5 million words identified 570 words of high frequency across a broad range of disciplines. So, the list can be a yardstick for validating further corpus studies, especially smaller ones. For this purpose and to gain further insight the ALAA list was compared with the AWL. 543 words were shared. Only 27 words were unique to the AWL. Table 6 shows the 27 AWL words which do not occur in ALAAs.

Table 6. The AWL words not found in the ALAA corpus

administrate	depress	levy	revenue
append	distort	nuclear	rigid
behalf	erode	offset	subsidy
cease	estate	prohibit	sum
collapse	export	purchase	suspend
consent	incentive	restore	transit
convene	injure	restrain	

These 27 words may hint at potential areas of bias in the AWL. For example, such words as *purchase, subsidy, estate, transit and depress* (which apparently features in the AWL thanks to depression) are strongly associated with economics.

However, the majority of AWL words feature in ALAAs, a fact which strengthens the observation that the AWL is generally robust and vastly permeates academic disciplines (Coxhead, 2011). This may also be related to the fact that applied linguistics is a multidisciplinary field and includes knowledge of multiple domains including society, psychology, language, and research.

However, occurrence is one thing and frequent occurrence is another. The ideal comparison of the frequency of AWL words in the original 3.5 million words and the present corpus would be to juxtapose the two frequency rankings. However, because the exact frequencies of AWL words were not available to the author and for the sake of convenience, the words in the first AWL sublist, the most frequent sublist in academic texts, according to Coxhead (2000), were compared with their rankings and frequencies in ALAAs (Tables 7A & B). Table 7A lists the first 60 words in the AWL with their ranks and frequencies in ALAAs. Table 7B gives the ranks and frequencies of the top 30 ALAA-shared AWL words and designates how they are distributed across the sublists. Again, confirmations and differences emerged. The fact that *analyze*, *approach*, *assessment context, data, process* and *research* rank very high in ALAAs establishes their strategic role. There are many words from AWL Sublist One whose frequencies are very low. Among these words are those which are usually associated with particular disciplines, e.g. *estimate, export, finance, income, labor, percent, legislate, sector*. Further analysis can reveal the exact ALAA status of words in other sublists. However, instead of tabulating the status of other sublists in ALAAs, the distribution of the 30 most frequent AWL words in ALAAs across the sublists is tabulated in Table 7B. Fourteen (46.66%) words belong to the first sublist, 8 (26.66%) to Sublist Two, and the rest (26.66%) to the next four sublists. In general, this evidences a strong academic identity for ALAAs as most of the words they share with the AWL hold a top position in terms of frequency of use.

Table 7A. The ranks and frequencies of the words from the first AWL sublist in ALAAs

AWL Words	Frequency in ALAAs	Rank In ALAAs	AWL Words	Frequency in ALAAs	Rank In ALAAs
Analyze	1345	30	individual	280	224
Approach	590	91	Interpret	249	248
Area	238	255	Involve	323	194
assessment	782	60	Issue	360	171
Assume	100	559	Labor	4	2959
Authority	20	1365	legal	24	1261
available	94	583	legislate	8	2135
benefit	169	373	major	157	389
concept	188	327	method	388	154
consistent	544	106	occur	158	390
constitute	71	708	percent	24	1268
context	811	56	period	123	479
contract	4	2856	policy	182	337
create	287	219	principle	145	421
data	705	72	proceed	111	520
define	123	477	process	911	51
derive	45	908	require	191	322
distribution	79	652	research	1723	26
economic	51	854	respond	90	610
environment	192	315	role	424	138
establish	122	482	section	121	487
estimate	38	1003	sector	8	2173
evidence	313	203	significant	466	120
export	0	-	similar	210	293
Factor	355	177	source	126	470
Finance	8	2108	specific	443	130
Formula	70	713	structure	391	152
Function	281	222	theory	518	103
Identify	625	85	variable	452	128
Income	1	6077			
Indicate	422	139			

For some AWL words, e.g. *proceed*, *academy*, the root words are not so frequent but the derived forms are. Some related AWL words, e.g. *specific* and *specify*, are listed separately while this procedure is not followed consistently in similar cases, e.g. *law* and *legal*.

Table 7B. The ranks and frequencies of the top 30 ALAA-shared AWL words and their distribution in the AWL sublists

AWL-ALAA Word	Frequency in ALAAs	Rank in ALAAs	AWL Sublist	AWL-ALAA Word	Frequency in ALAAs	Rank in ALAAs	AWL Sublist
Research	1723	26	1	Academic	538	100	5
Analyze	1345	30	1	Theory	518	103	1
Task	1151	37	3	Significant	466	120	1
Instruct	926	48	6	Specific	443	130	1
Process	911	51	1	Role	424	138	1
Context	811	56	1	Complex	394	146	2
Assess	782	60	1	Feature	373	160	2
participate	766	63	2	implicate*	368	166	4
Interact	764	64	3	Design	362	169	2
investigate	726	68	4	Issue	360	171	1
Data	705	72	1	Factor	355	177	1
Text	699	74	2	Target	315	201	5
Strategy	653	81	2	Evidence	313	213	1
Focus	627	84	2	Aspect	293	214	2
Approach	590	91	1	Accurate	291	217	6

*It should be noted that in ALAAs, *implication*, which is the source of high frequency for *implicate*, is more related to *imply*, which does not figure in this table, than *implicate*.

Checking the AWL words against the list of ALAA keywords is also revealing. While only 95 (7.12%) of the 1,334 GSL words in the ALAAs corpus are frequent enough to be key (Table 5), 132 (24.30%) of 543 AWL words occurring in ALAAs feature as keywords (Table 8). Apart from the 27 AWL words which do not occur in ALAAs, 410 AWL words do not occur frequently enough to be a key. The share of the AWL in the 375 key word families turns out to be 35.2%, in some contrast to a share of 25.53% for GSL words.

Table 8. The top 132 AWL words occurring as ALAA Keywords

analyze	implement	academy	construct
assess	instruct	access	context
focus	interact	coherent	culture
identify	investigate	communicate	domain
motive	participate	complex	dynamic
highlight	research	compute	edit
hypothesis	text	constrain	emerge

empirical	challenge	grade	overall
enhance	clause	ideology	paradigm
evaluate	code	image	parameter
facilitate	complement	implicate	perceive
impact	concept	implicit	positive
integrate	conclude	incorporate	potential
perspective	consult	input	prime
practitioner	converse	insight	professional
predict	create	institute	protocol
process	criterion	intense	random
qualitative	data	intermediate	rely
relevant	debate	interpret	resource
revise	demonstrate	intrinsic	reveal
task	diverse	involve	sequence
theory	draft	journal	significant
topic	emphasis	label	similar
vary	ensure	major	simulate
accurate	environment	manipulate	statistic
acquire	evolve	maximize	strategy
adult	expert	media	survey
approach	explicit	method	target
appropriate	format	mode	transfer
assist	framework	modify	transform
automate	gender	monitor	underlie
aware	generate	orient	valid
benefit	globe	outcome	violate

All this said, the account of the lexical make-up of ALAAs presented above should be treated cautiously and as preliminary indications. The broad hints about the state of lexis or particular lexical items given by the output of the analyses here may need modification when considered in context. Words are not monolithic units. They have multiple and context-bound meanings and nuances of meaning and interact with other elements in the process of use. For example, *manipulate*, which is used polysemously in general phraseology, almost constantly occurs with neutral or moderately positive connotations in ALAAs to refer to, say, vocabulary or tool use. While, out of 29 occurrences, *manipulate* is used negatively only once-- in adjectival form in an abstract about critical literacy and the importance of raising consciousness to the "manipulative power of text" over people-- the opposite may be the case in political discourse. So, if we want a clear and comprehensive understanding of the differential frequencies of ALAA

words, we need to go beyond this preliminary step and take into account both their linguistic context and the socio-cultural context which gives rise to them (Kress, 1989).

7. Conclusion

This study for the most part provides a quantitative sketch of the stock of words which have been used in the abstracts of ALAAs published in recent years. This achievement seems noteworthy because lexis plays a major defining role in both formal and ideational features of genres. There are serious shortcomings in this study which keep it short of offering a clear and comprehensive picture of the lexical make-up of the abstracts in this academic area. The corpus was too small to allow final claims. A corpus of more than one million words may again give similar patterns of word choice and frequency; but then one could make more confident conclusions. The interpretation of the lexical output here is based on frequencies without any consideration of the behavior of the items in specific contexts, but a more qualitative approach could certainly provide deeper insights. Most words have multiple meanings; but this fact is skated over in different analyses reported in this paper. Even with a purely quantitative approach one could be more fine-tuned and focus on the terms used to report on specific areas of inquiry. Keywords analysis helps but it suffers from two serious flaws: compound words are stripped to their components while many key notions are communicated through compounds; and, the types of the same lemmas are reported separately making it difficult to have a coherent picture of keywords. Finally, although having some information about frequent words and comparing them with well-known lists are very useful, there may also be less frequent words which are of defining significance to ALAAs.

Still, the tables and statistics offered in this report can bestow a preliminary, but telling, portrayal of ALAAs or feed into the schemas of those already initiated in this subgenre and lift their bird's eye view up to an eagle's one. The choice of words and the frequency with which they are used tell a lot about the overriding processes in applied linguistics and the ideas current in it. ALAAs' main goal is to report in a condense way research about language learning, teaching, assessment, policy, etc. So, it is expected that words like *instruct*, *learning*, *context*, *process* are frequent in them. They also include frequent research-specific words, such as *theory*, *approach*, *data*, *focus*, *design*. One more achievement of this research can be exposing the degrees of association of ALAAs and their lexical ingredients with general and mainstream academic texts by specifying the frequency of ALAA words and their standing in the AWL and the GSL. This study and the words listed here can be an aid in teaching the vocabulary of the field

and helping students develop their academic reading and writing ability. Teachers can use the lists as points of departure in preparing materials including examples in which these words are used or employ concordance tools to show the words in context and examine their linguistic features and grammatical behavior.

The research can also generate more fine-tuned questions, e.g. the frequent collocations or chunks in ALAAs and serve as an introduction to the exploration of the lexical profile of the main body of those articles. Studying the changes which have occurred through the years in the lexical make-up and frequency of the words used in ALAAs or a comparative study of lexical use and frequency in the abstracts produced by native and non-native writers can also be of interest to applied linguists and others.

References

van Aalst, J. (2010). Using Google Scholar to estimate the impact of journal articles in education. *Educational Researcher*, *39*(5) 387-400.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6 Ed.). Washington, D.C.: Author.

Bauman, J. & Culligan, B. (1995). General service list. Retrieved, July 21, 2012 from the World Wide Web: http://plaza3.mbn.or.jp/~bauman/gsl.html.

Bhatia, V. K. (1993). *Analyzing genre: Language use in professional settings*. London: Longman.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: ttp://www.natcorp.ox.ac.uk/

Busch-Lauer, I.-A. (1995). Abstracts in German medical journals: A linguistic analysis. *Information Processing and Management*, *31*(5), 769–76.

Chan, S & Foo, S.K. (2004). Interdisciplinary perspectives on abstracts for information retrieval. *Iberica, 8*, 101-124.

Cobb, T.M. (2011). *Compleat Lexical Tutor*. Retrieved December 10, 2011 from the World Wide Web: http://www.lextutor.ca/

Coxhead, A. (1998). *An academic word list*. Wellington: Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2) 213-238.

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, *42*(2), 355-362.

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. *Journal of Documentation*, *62*(4), 428-446.

Flowerdew, J. (2001). Concordancing as a tool in course design. In M. Ghadessy, A. Henry & R.L. Roseberry (Eds.). *Small corpus study and ELT* (pp. 71-92). Amsterdam: John Benjamin.

Frances, W.N. & Kucera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin.

Ghadessy, M. (1999). Thematic organization in academic article abstracts. *Estudios Lingleses de la Universidad Complutense*, *7*, 141-161.

Groom, N. (2005). Patterns and meaning across genres and disciplines: An explanatory study. *Journal of English for Academic Purposes*, *4*(3), 257-277.

Hartley, J. (2003). Improving the clarity of journal abstracts in psychology: The case for structure. *Science Communication*, *24*(3), 366-379.

Hartley, J., & Benjamin, M. (1998). An evaluation of structured abstracts in journals published by the British psychological society. *British Journal of Educational Psychology*, *68*, 443-456.

Hartley, J., & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading*, *20*(2), 122-136.

Huckin, T. (2001). Abstracting from abstracts. In M. Hewings (Ed.). *Academic Writing in Context* (pp. 93-103). Birmingham: University of Birmingham Press.

Hyland, K. (2004). *Disciplinary discourses: Social interaction in academic writing*. Ann Arbor: University of Michigan Press.

Kress, G. (1989). *Linguistic processes in sociocultural practice*. Oxford: Oxford University Press.

Lorés, R. (2004). On RA abstracts: From rhetorical structure to thematic organization. *English for Specific Purposes*, *23*(3), 280-302.

McEney, M. & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Martin, P. M. (2003). A genre analysis of English and Spanish research paper abstracts in experimental social sciences. *English for Specific Purposes*, *22*(1), 25-43.

Nation, I.S.P. (2001). Learning vocabulary in another language. Cambridge : Cambridge University Press.

Pho, P. Z. (2008). Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. *Discourse Studies*, *10*(2), 231-250.

Praninskas, J. (1972). *American university word list*. London: Longman.

Romer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(1), 140-162.

Salager-Meyer, F. (1990). Discourse flaws in medical English abstracts: A genre analysis per research and text type. *Text*, *10*(4), 365-384.

Samraj, B. (2002). Disciplinary variation in abstracts: The case of wildlife behavior and conservation biology. In J. Flowerdew (Ed.). *Academic discourse* (pp. 40-56). Harlow; New York: Longman.

Samraj, B. (2005). An exploration of genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, *24*, 141-156.

Santos, M. B. (1996). The textual organization of research paper abstracts in applied linguistics. *Text*, *16*(4), 481-499.

Simpson-Vlach, R. & Ellis, N. (2010). An academic formula list: New methods in phraseology research. *Applied Linguistics*, *31*(4) 487-512.

Stubbs, M. (2010). Three concepts of keyness. In M. Bondi & M Scott (Eds.). *Keyness in texts* (pp. 21-43). Amsterdam: John Benjamin.

Swales, J.M. & Feak C.B. (2009). *Abstracts and writing of abstracts.* Ann Arbor: University of Michigan Press.

Vongpumivitch, V., Huang, J.-y. & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *Enlish for Specific Purposes*, *28*(1), 33-41.

West, M. (1953). *A general service list of English words*. Longman: London.

Ventola, E. (1994). Abstracts as an object of linguistic study, in S. Cmejrkova, F. Danes & E. Havlova (Eds.). *Writing vs. speaking: Language, text, discourse, communication*. Proceedings of the conference held at the Czech Language Institute of the Academy of Sciences of the Czech Republic, Prague, 14–16 October, 1992 (pp. 333–52). Tubingen: G. Narr.

Xue, G. & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication*, *3*, 215-229.

Appendix
Applied linguistics journals from which abstracts were taken, with the number of abstracts and abstract words used in the study given in parentheses

1. *Annual Review of Applied Linguistics* (87; 15,590)
2. *Applied Linguistics* (145; 25,712)
3. *ELT Journal* (198; 26,576)
4. *Language Teaching Research* (126; 23,851)
5. *Language Testing* (119; 23,590)
6. *Modern Language Journal* (176; 31,665)
7. *English for Specific Purposes* (130; 24,869)
8. *English Teaching_ Practice & Critique* (147; 25,116)
9. *Foreign Language Annals* (172; 29,826)
10. *Journal of Second Language Writing* (92; 15,616)
11. *Language Learning* (171; 31,555)
12. *Language Learning & Technology* (71; 13,496)
13. *Second Language Research* (105; 20,844)
14. *System* (212; 40,433)
15. *TESOL Quarterly* (120; 25,422)